Study of Fuel Quality in the State of Maranhão Through Principal Component Analysis

Morgana Cristhya Silva dos Santos[1*], Luciana Pereira Barbosa[1], Allan Kardec Duailibe Barros Filho[1]
[1]Postgraduate Program in Electrical Engineering, Federal University of Maranhão; São Luis, Maranhão, Brazil

This study evaluated the quality of S10 and S500 diesel sold in the state of Maranhão, using data from the National Petroleum Agency (ANP) and the Principal Component Analysis (PCA) technique, to identify patterns in quality indicators. PCA simplified the data, highlighting the most relevant characteristics that influence diesel quality. The results revealed relationships between indicators and groups of samples with similar characteristics, which can be helpful for fuel monitoring. The research demonstrated that PCA is an effective tool to assist in the assessment and control of fuel quality, highlighting the importance of continuous monitoring and advanced statistical analysis in determining fuel quality.
Keywords: PCA. Quality. Fuels.

In Brazil, fuels are produced in several different regions and basins. The quality and properties of these hydrocarbons depend on their region of origin. However, these characteristics are also subject to change at the retail stations. Currently, there are approximately 43,266 petroleum product resale stations distributed throughout the country [1], which undergo regular monitoring and inspection processes throughout the year. To this end, the National Petroleum and Biofuels Agency (ANP) set up the Fuel Quality Monitoring Program (PMQC) in 1998, which is used to identify areas of fuel non-compliance. The analysis of these samples refers to various technical requirements established by the program. The fuels most consumed in Brazil are gasoline and diesel [2,3]. Diesel is the primary fuel used to transport passengers, cargo, and agricultural products in Brazil.

In February 2024, diesel production in Brazil increased by approximately 8.5% compared to the previous year and continues to exceed growth expectations for the current year [4]. Maranhão reflects this national diesel consumption, mainly due to intense agricultural activities and cargo transportation. In terms of characteristics, diesel has chains composed of 8 to 16 carbons and has lower concentrations of nitrogen, sulfur, and oxygen [3]; however, these characteristics can be transformed with the addition of other substances. Fuel contamination can cause several problems in terms of burning and storage quality, the latter of which is directly related to the oxidative stability of fuels, referring to how well they resist degradation processes [5]. Additionally, interference in fuel composition is a concern, primarily due to engine operation and the release of atmospheric pollutants. In this regard, the monitoring and control of these fuels are critical, as indicated by the compliance indices (%IC) observed in the PMQC [6].

Identifying these parameters can ensure that the fuels that reach the consumer are increasingly better. Monitoring is now carried out in most of Brazil, in partnership with educational and research institutions. According to the ANP Statistical Yearbook [7], in 2020, Maranhão had 1,477 fuel retail outlets. The following year, it monitored 170 municipalities, with 192 samples of ethanol, 1,035 of gasoline, and 1,021 of diesel oil.

In Maranhão, fuel quality analysis is carried out by the Laboratory of Analysis and Research in Petroleum Analytical Chemistry at the Federal University of Maranhão (LAPQAP/UFMA). The laboratory's objectives are to automate the analysis and data processing processes as a decision support tool. A single sample offers a set of variables [8],

Received on 18 May 2025; revised 25 July 2025.
Address for correspondence: Morgana Cristhya Silva dos Santos. Avenida Hiram Saboia, 777, Centro, Balsas. São Luis, Maranhão. Zipcode: 65800-000. E-mail: morgana.santos@discente.ufma.br.

J Bioeng. Tech. Health 2025;8(4):333-338

so the acquisition of fuel samples from Maranhão, collected together with all their physicochemical properties, constitutes a high-dimensional database, making it difficult to process them. In this sense, the main characteristics analyzed for diesel types are Distillation (10% and 50%), Specific Mass at 20°C, Biodiesel Content, Boiling Point, and Color.

Advances in computational techniques for data analysis are becoming increasingly important for classification. Thus, a common technique is Principal Component Analysis (PCA) [9]. PCA is a technique that has been applied to the treatment of multivariate data and has yielded satisfactory results [8]. As it is an exploratory technique, it enables the identification of correlations between quality indices, the analysis of irregular samples, and the examination of relationships between measured variables, as well as the identification of relationships or groupings within samples. This type of analysis can offer a more efficient arrangement of data distribution in a smaller set than the original, while preserving most of the information.

Therefore, this paper proposes the application of Principal Component Analysis for the exploratory analysis of fuel quality data, aiming to reduce data dimensionality and identify possible patterns and correlations between fuel quality indicators in the state of Maranhão.

Principal Component Analysis (PCA)

Principal Component Analysis applied to the analysis of data with a large number of variables can be described as presented by Correa [10] as a method that evaluates interrelationships with the "*aim of recognizing patterns in the distribution of samples, evaluating the relationship between samples and variables, and also detecting the presence of samples that show a distinct behaviour (outliers)*". PCA is a dimensionality reduction technique that transforms a data set with a large number of variables into a smaller set. Data reduction using PCA is achieved by linearly combining the correlations of the original variables. In this way, a smaller representation is obtained through the resulting principal components (PCs). The mathematical model behind PCA can be described as follows, where the matrix X is decomposed according to Equation 1 [10, 11].

$$X_{nxm} = U_{nxn}\Sigma_{nxm}V^{T}_{mxm} \quad (1)$$

Where n is the number of samples and m is the number of initial attributes that make up the data set of interest, X. The PCA assumption is that the first component should have the maximum variance explained and the second the variance not explained in the first component [10]. While Unxn e Vmxm are orthogonal and Σ is a diagonal matrix made up of the singular values [11]. Moreover, P is the weight matrix in which the elements in each column correspond to the coefficients of the linear combinations of the original variables, as shown in Equation 2 [10].

The results of applying PCA can be visualized using score plots, where the main relationships between the variables are clearly visible. With the different groupings of samples, similarities and differences, as well as trends and outliers, are identified.

Equation 2.

$$P = X.V^{T} = X_{nxm}.\begin{bmatrix} | & | & | & | \\ v_1 & v_2 & \cdots & v_p \\ | & | & | & | \end{bmatrix} = \begin{bmatrix} PC_{1,1} & PC_{1,2} & \cdots & PC_{1,q} \\ PC_{2,1} & PC_{2,2} & \cdots & PC_{2,q} \\ \vdots & \vdots & \ddots & \vdots \\ PC_{n,1} & PC_{n,2} & \cdots & PC_{n,q} \end{bmatrix}$$

Materials and Methods

To conduct this research, the state of the art on the subject was initially surveyed. After this, data on fuel quality in the state of Maranhão were acquired from the ANP's open database [12]. This data is obtained through the ANP's Fuel Quality Monitoring Program (PMQC).

The PMQC provides data on the analysis of the quality of diesel oil, hydrated ethanol, and gasoline. For this research, data were collected through the monitoring of the quality of ordinary diesel fuel (S10 and S500), with samples taken from various fuel stations in different municipalities in Maranhão throughout 2023.

After selecting the data, it was processed in accordance with the objectives of this work. Samples of ordinary S10 and S500 diesel oil and the tests corresponding to these samples were selected. These tests make up the group of variables used to apply the PCA technique. Figure 1 shows the characteristics analyzed by PCA in this study.

In Figure 1, the diagram shows the fuels analyzed and their respective tests. Samples with null data for the tests were excluded from the study. For the ordinary diesel oil samples selected, the following tests were considered as variables for the application of PCA: (1) distillation - 10%, (2) distillation - 50%, (3) specific mass at 20°C, (4) biodiesel content, (5) flash point, and (6) color. As the result of the color test is found in the data set as a word, rather than a number, for the PCA application, these results were replaced with numerical values (Table 1).

Table 1. Numerical correspondents assigned to the colors of the samples for the application of PCA in Matlab.

Sample Color	Corresponding Number
Yellow	1
Orange	2
Red	3

After selecting the characteristics to be analyzed using PCA, a data matrix was assembled, with rows containing samples of standard diesel oil (S10 and S500) and columns containing the variables corresponding to these samples, forming a 775 × 7 matrix. With the matrix formed, an algorithm was implemented in MATLAB R2023b software (academic version) to apply PCA to the study dataset.

Results and Discussion

The application of the PCA technique to the data sets related to fuel quality in the state of Maranhão, as provided by the PMQC bulletins available on the ANP website, enabled the acquisition of results in the form of graphs and tables relating to the application of this technique to the data under study.

Table 2 presents the percentage of explained and accumulated variance values, as well as the number of principal components (PCs) obtained by applying the PCA method to the data for the S10 and S500 ordinary diesel fuel samples.

Figure 1. Diagram illustrating the characteristics analyzed by the PCA technique in this research.

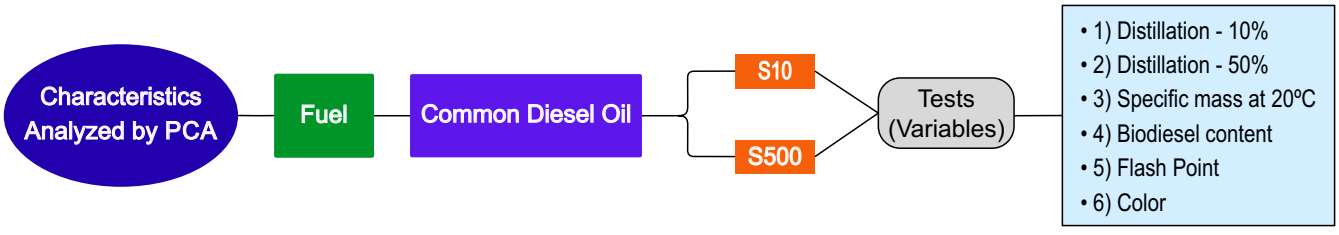


Table 2. Values, in percentages, of explained variance and accumulated variance obtained by applying PCA to the data referring to S10 regular diesel oils and S500 regular diesel oils.

PC	Explained Variance (%)	Accumulated Variance (%)
1	36.68	36.68
2	18.25	55.63
3	14.53	70.16
4	13.18	83.34
5	11.26	94.61
6	5.39	100.00

According to Table 2, the 6 PCs obtained after applying PCA to the diesel oil samples are observed. Therefore, this number of PCs was modeled in the algorithm implemented in Matlab. One of the objectives of applying the PCA technique is to reduce the dataset's dimensionality without losing information. Therefore, when applying this technique, the result must show that the maximum number of PCs needed to represent the data set is lower than the number of original variables considered in the analysis [13]. Therefore, in this work, the application of PCA made it possible to reduce the number of variables analyzed, eliminating the PCs that presented lower explained variance. In this study, the explained and accumulated variances represent how the chemical information distributed in the original variables can be represented by a smaller number of variables, which are the PCs.

Thus, it was found that it is possible to eliminate PC6, which has the lowest explained variance (5.39%), and still have more than 94% of the data explained by the set of data analyzed with only the five remaining PCs, leading to a reduction in the dimensionality of these data. Furthermore, it is observed that the PC's that most contribute to the representation of the information contained in the common diesel oil data are the first four PC's, which together represent more than 80% of the total variance of these data, with PC1 and PC2 being the ones that contribute most to this representation.

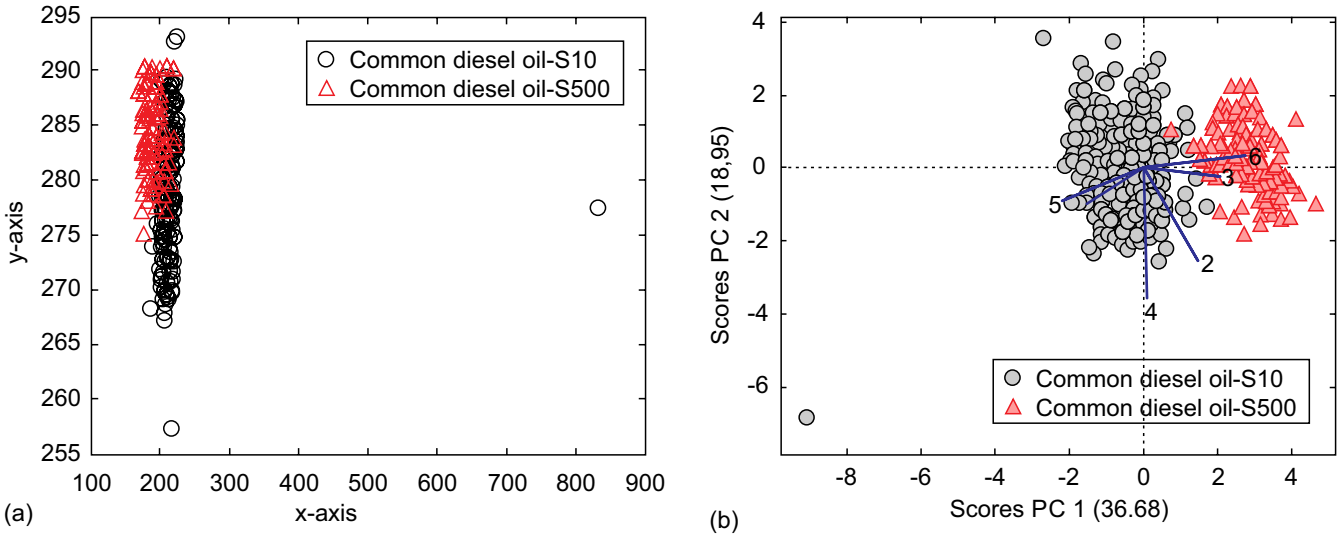
Figure 2 shows graphs with data from diesel oil samples before applying PCA (see Figure 2 (a)) and a biplot graph of scores and loadings of PC1 in relation to PC2 after applying PCA (see Figure 2 (B)).

When comparing Figure 2(a) with Figure 2(b), it is observed that the application of PCA allows for a better visualization of the data related to the diesel oil samples. Furthermore, Figure 2(b) shows that after applying PCA, a grouping was obtained between data from samples of the same type and a separation between data referring to samples of standard diesel oil S10 and data referring to diesel oil samples standard S500.

The data for standard diesel oil S10, which is further removed from the vast majority of data (see Figures 2(a) and 2(b)), corresponds to the samples that, in the tests, exhibited an orange color, a distinct color from the others observed. Because, in the other samples, the S10 diesel oil presented a yellow color in the tests. S500 diesel oil was red in the tests.

Thus, it is observed that color was a determining variable for both the grouping and separation of data for standard diesel oil S10 and S500. Figure 2(b) presents data relating to diesel oil samples in the biplot graph of scores and loandings obtained after applying PCA, in which each symbol represents a sample of standard diesel oil S10 or S500 in relation to PC1 and PC2, which are the scores. The blue lines represent the weights of the analyzed variables (tests referring to these samples, which PMQC carried out) in relation to PC1 and PC2, which are the loandings. When analyzing the biplot graph, it is essential to consider that the relevance of the variables analyzed in this study can be measured by the size and direction of loadings (blue lines on the biplot graph). In this case, the biodiesel content variable has a greater weight in relation to the PC2 axis. On the other hand, the color variable, represented

Figure 2. Graphs with data from standard diesel oil samples S10 and S500: (a) graph with data before applying PCA; (b) biplot graph of scores and loandings with the data after applying PCA.



in Figure 2(b) by line 6, has a greater weight in relation to the PC1 axis.

The PCs are the new variables obtained through the application of PCA, and they are capable of explaining a certain amount of characteristics of the analyzed dataset, in this case, the samples of standard diesel oil S10 and S500. According to the explained and accumulated variances presented in Table 2, it is observed that PC1 and PC2 explain over 50% of the data set analyzed in this study. Thus, the variables biodiesel content and color contribute more to explaining the variations present in the data from the standard diesel oil samples S10 and S500.

Analyzing Figure 2(b), it is possible to observe correlations between the variables being analyzed. As the straight lines represent the original variables analyzed, it is observed that there is a correlation between the variables biodiesel content (4) and flash point (5) in relation to the color variable (6). This correlation can be observed, as the straight lines have an angle of almost 180° between them, indicating that the variables related to these straight lines are correlated and inversely proportional to each other [13].

According to the results obtained, it is noted that research carried out on the quality of S10 and S500 diesel in the state of Maranhão is of great relevance both for the state, where the fuel sector plays a crucial role in energy supply, and for assessing fuel quality, which is essential to ensure engine efficiency, reduce atmospheric emissions, and protect public health. The application of Principal Component Analysis (PCA) enables you to monitor diesel quality more precisely, identifying patterns and trends that help classify and control the fuel, thereby guaranteeing that end consumers receive quality fuel.

Conclusion

The use of the Principal Component Analysis technique is relevant when we want to simplify the analysis of large datasets. In this study, the application of the PCA technique to data from samples of standard diesel oil (S10 and S500) allowed for an exploratory analysis of several variables that contribute to determining the quality of this fuel. We observed that applying PCA to data from diesel oil samples enabled a reduction in the dataset's dimensionality without loss of information by extracting the main components. Therefore, the first four PCs are capable of representing the variance of more than 80% of the studied dataset.

Furthermore, the color variable was highly relevant for analyzing these results, as it is a crucial factor in determining the groupings and separations between the analyzed diesel oil data, in addition to correlating with other variables studied, such as biodiesel content. Therefore, color can help in determining possible patterns existing between the characteristics of the data referring to samples of standard diesel oil S10 and S500.

Therefore, the application of PCA provided both data dimensionality reduction and the identification of groupings and correlations between the analyzed variables, thus allowing the objective of this work to be achieved. This demonstrates that this technique can contribute to the ongoing improvement of programs such as the ANP's PMQC, thereby enhancing the quality of fuels sold at stations not only in the state of Maranhão but also throughout Brazil.

For future work, it is proposed to apply the Principal Component Analysis technique to other fuels, such as gasoline and ethanol, to verify the accuracy of this technique in these additional fuels and in a larger dataset.

References

1. Brasil. Anuário Estatístico Brasileiro do Petróleo, Gás Natural e Biocombustíveis: 2023. Rio de Janeiro: Agência Nacional do Petróleo, Gás Natural e Biocombustíveis; 2006.
2. Oliveira EP, et al. Investigação do teor de água no biodiesel utilizado na composição do diesel B comercializado por uma distribuidora de combustíveis em Manaus/AM. Braz J Dev. 2021;7(9):89663-80.
3. Ribeiro CB, Schirmer WN. Panorama dos combustíveis e biocombustíveis no Brasil e as emissões gasosas decorrentes do uso da gasolina/etanol. BIOFIX Sci J. 2017;2(2).
4. Instituto Brasileiro de Petróleo (IBP). Boletim do ciclo diesel 2024 [Internet]. Rio de Janeiro: IBP; 2024. Available from: https://www.ibp.org.br/personalizado/uploads/2024/03/boletim-ciclo-diesel-i-marco-de-2024-9.pdf
5. Moura HO, et al. Advances in chemometric control of commercial diesel adulteration by kerosene using IR spectroscopy. Anal Bioanal Chem. 2019;411:2301-15.
6. Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP). Boletim de biocombustíveis e qualidade de produtos 2022b [Internet]. ANP; 2022. Available from: https://www.gov.br/anp/pt-br/centrais-de-conteudo/publicacoes/boletins-anp/arquivos-boletim-de-biocombustiveis-e-qualidade-de-produtos/boletimsbq2022.pdf
7. Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP). Anuário Estatístico Brasileiro do Petróleo, Gás Natural e Biocombustíveis 2021a [Internet]. ANP; 2021. Available from: https://www.gov.br/anp/pt-br/centrais-de-conteudo/publicacoes/anuario-estatistico/arquivos-anuario-estatistico-2021/anuario-2021.pdf
8. Blanco ALP, Carauta ANM. Análise de componentes principais aplicada à espectroscopia no infravermelho de misturas de diesel e biodiesel: estudos de casos. Rev Souza Marques. 2018;18(37):9-44.
9. Alves WF, et al. Análise multivariada dos parâmetros físico-químicos da gasolina "tipo C" comercializada no Vale do Juruá-Acre. South Am J Basic Educ Tech Technol. 2019;6(1).
10. Correa C. Metodologias analíticas para avaliar a biodegradabilidade do diesel, biodiesel e blendas B10 [tese]. Porto Alegre: Universidade Federal do Rio Grande do Sul; 2021.
11. Parente LER. Análise exploratória de perfilagem acústica para avaliação da qualidade de cimento com simulações computacionais. Rio de Janeiro: Departamento de Engenharia Mecânica, Pontifícia Universidade Católica; 2022.
12. Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP). PMQC – programa de monitoramento da qualidade dos combustíveis 2023c [Internet]. ANP; 2023. Available from: https://www.gov.br/anp/pt-br/centrais-de-conteudo/dados-abertos/pmqc-programa-de-monitoramento-da-qualidade-dos-combustiveis
13. Folli G, et al. Tutorial para aplicação didática de quimiometria em software gratuito – Parte I: análise de componentes principais em dados de infravermelho médio e propriedades físico-químicas de amostras de petróleo. Rev Ifes Ciênc. 2023;9(1):1-14.